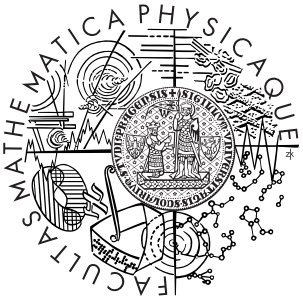
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Petr Hanek

Bradley-Terry model

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: RNDr. Šárka Došlá, Ph.D.
Studijní program: Obecná matematika

2010

Děkuji vedoucí své bakalářské práce RNDr. Šárce Došlé, Ph.D., za cenné rady a náměty, odborné vedení a trpělivost během tvorby mé práce. Dále bych chtěl poděkovat svým přátelům za shovívavost, a pracovním kolegům, kteří mi umožnili svědomitě pracovat na mé práci.

Prohlašuji, že jsem svou bakalářskou práci napsal samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce a jejím zveřejňováním.

V Praze dne 5.srpna 2010 Petr Hanek

Obsah

Název práce: Bradley-Terry model
Autor: Petr Hanek
Katedra (ústav): Katedra pravděpodobnosti a matematické statistiky
Vedoucí bakalářské práce: RNDr. Šárka Došlá, Ph.D.
e-mail vedoucího: dosla@karlin.mff.cuni.cz

Abstrakt: Tato bakalářská práce se zabývá Bradley-Terryho modelem a jeho aplikacemi. Bradley-Terryho model patří mezi logistické modely pro párová porovnávání. V textu je popsán zobecněný lineární model a modely logistické regrese a odvození základního Bradley-Terryho modelu. Dále se zabýváme jeho rozšířeními pro různé situace a metodami pro odhad jeho parametrů. Součástí práce je popis funkcí statistického balíčku v programu R pro práci s tímto modelem. Důležitou částí práce je pak aplikace nastudovaných metod na data z české volejbalové ligy, kde se zabýváme odhadem vlivu různých parametrů sportovního týmu na jeho úspěšnost v utkáních.

Klíčová slova: logistický model, Bradley-Terryho model, párové porovnávání

Title: Bradley-Terry model
Author: Petr Hanek
Department: Department of probability and mathematical statistics
Supervisor: RNDr. Šárka Došlá, Ph.D.
Supervisor's e-mail address: dosla@karlin.mff.cuni.cz

Abstract: We study the Bradley-Terry model and its applications. The Bradley-Terry model belongs to the class of the logistic models for pairwise comparisons. First, we describe the generalized linear model and the logistic regression model. Then the basic Bradley-Terry model is derived. Furthermore, we deal with some extensions of the basic model and with methods for estimation of its parameters. The thesis also contains a description of a statistical package for the program R. An important part of the work is an application of the studied methods on real data from the Czech volleyball league. Here we estimate the impact of different parameters of the sports teams on their probability of success in the matches.
Keyword: logistic model, Bradley-Terry model, pairwise comparison

Úvod

Představte si situaci. Je před vás položen soubor věcí, skupina individuí, nějaké množství objektů. Ty mohou být libovolné. Jistou vlastnost ovšem musíme předpokládat. Pro tato tělesa je charakteristické, že existuje určitý aspekt, podle kterého se dají objekty porovnávat. Jde tedy o nějakou kvantitativní vlastnost. Vaším úkolem je sestavit pořadí prvků souboru založené na již zmíněné charakteristice.
Jednou z možností, jak k této problematice přistoupit, je vzájemné porovnávání výsledků. Toto srovnávání je založeno na principu párového porovnávání. Z daného souboru vybereme vždy dvojici. Z této dvojice určíme vítěze. Touto metodou budeme schopni sestavit pořadí všech prvků.
Co si pod souborem prvků nebo množinou jedinců představit? Téměř cokoliv, co nás napadne. Máte několik druhů sýrů. Vzájemným srovnáním získáte spektrum sýrů od nejchutnějšího po ten nejméně lahodný. De facto jakýkoli vzorek potravin. Metoda je uplatitelná i při degustaci vín. Z jiného soudku. Máte skupinu závodníků, sportovců, týmů. Vzájemnými výsledky mezi každou dvojicí opět obdržíte jejich pořadí. Ve výčtu možností, co si pod sledovaným souborem představit, bychom mohli pokračovat dále.

Bradley-Terryho model je jeden z nepřeberného množství statistických modelů. Je pojmenován po dvou význačných matematicích 20.století — Allanu Bradleym a Milton E. Terrym. Ti jej světu představili v roce 1952.
Model je založen na principu párového porovnávání. Mějme několik objektů. Pro lepší představu si vezměme hráče tenisu. Tito hráči mezi sebou odehrají obecně různý počet utkání. Dle počtu vítězných utkání jsme schopni určit lepšího z dané dvojice. Pro tuto dvojici již tedy máme určeno pořadí. Vzájemným srovnáním všech možných dvojic z daného souboru získáme celkové pořadí jednotlivců. Tato aplikace je možná prakticky pro veškerá sportovní klání a soutěže. Uplatnění nachází rovněž i v matematické statistice, biogenetice, gastronomii, při zkoumání chování živočichů v přírodě. Model nám dává odhad pravděpodobnosti vítězství jednoho týmu nad druhým, co se týče sportu, popřípadě určuje pořadí týmů.

Tato bakalářská práce je rozdělena do čtyř kapitol. V první kapitole je nejprve připomenut klasický lineární regresní model a zobecněný lineární model. V sekci 1.4 je pak popsán model logistické regrese pro binární data, odkud je již jen krůček k odvození základního Bradley-Terryho modelu.

Ve druhé kapitole se již zabýváme Bradley-Terryho modelem. Základní model je odvozen v sekci 2.1. V následující sekci 2.2 jsou pak popsány modifikace základního modelu pro různé situace, se kterými se můžeme při aplikacích (zejména na sport) setkat — Bradley-Terryho model s efektem domácího prostředí, Bradley-Terryho model pro charakteristiky hráčů, Bradley-Terryho model zahrnující remízy a další. V sekci 2.3 jsou potom uvedeny některé postupy pro numerické řešení a získání odhadů parametrů v představeném modelu.
Třetí kapitola se soustředí na práci s Bradley-Terryho modelem v prostředí statistického softwaru R. V této kapitole jsou popsány všechny důležité funkce balíčku `BradleyTerry`, všechny ilustrované na příkladě z tenisového prostředí, včetně uvedených výstupů z programu R. Upozorníme také na nedostatky tohoto balíčku a podáme krátký přehled o chystaném balíčku `BradleyTerry2`.
Poslední, čtvrtá kapitola, se zabývá statistickou analýzou výsledků volejbalových mužstev pomocí Bradley-Terryho modelu. Data o jedenácti volejbalových týmech ze tří sezon nejvyšší české soutěže analyzujeme pomocí modelů uvedených v předchozích kapitolách a snažíme se určit, jaké faktory charakterizující volejbalový tým mají dopad na soutěžní výkony mužstva.

Kapitola 1

Zobecněný lineární model a logistický model

Cílem regresních modelů je vysvětlit změny hodnot (variabilitu, kolísání) náhodného vektoru pomocí jedné či několika nenáhodných nezávislých proměnných. V modelu uvažujeme nekorelované náhodné veličiny $Y_1, Y_2, \ldots, Y_n$. Variabilitu náhodného vektoru $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ vysvětlujeme pomocí n vektorů známých hodnot $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}$. Veličiny $Y_1, Y_2, \ldots, Y_n$ nazýváme závislými, nebo vysvětlovanými proměnnými, někdy se také používá výraz odezvy. Vektory $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}$ označujeme jako nezávisle proměnné, prediktory, popřípadě vysvětlující proměnné.

1.1 Klasický lineární regresní model

Klasickým lineárním regresním modelem rozumíme model ve tvaru:

$$Y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_k x_{ik} + \epsilon_i, \qquad i = 1, 2, \ldots, n,$$

kde $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)'$ jsou neznámé parametry modelu a ϵ_i označuje náhodné složky.
Tento model lze přepsat maticově:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{12} & x_{13} & \dots & x_{1k} \\ 1 & x_{22} & x_{23} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \dots \\ 1 & x_{n2} & x_{n3} & \dots & x_{nk} \end{pmatrix}, \boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}, \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

V modelu se obvykle předpokládají následující vlastnosti:

- $\mathrm{E}(\epsilon_i) = 0, \qquad i = 1, \dots, n$
- $\mathrm{var}(\epsilon_i) = \sigma^2 < \infty, \qquad i = 1, \dots, n$
- $\mathrm{cov}(\epsilon_i, \epsilon_j) = 0, \qquad i, j = 1, \dots, n \quad i \neq j$
- $\mathrm{h}(\boldsymbol{X}) = k$
- ϵ_i má normální rozdělení $\qquad i = 1, \dots, n$

Podrobnější informace o klasickém lineárním modelu nalezneme například v [13].

1.2 Zobecněný lineární model

Zobecněné lineární modely se často značí zkratkou GLM (z anglického Generalized Linear Model). Předpokládáme, že $Y_1, Y_2, \dots, Y_n$ jsou nezávislé a patří do tzv. exponenciální rodiny pravděpodobnostních rozdělení, viz 1.3. Pak GLM rozumíme model pro než platí:

$$\mathrm{E}Y_i = g(\boldsymbol{x_i}'\boldsymbol{\beta}) \qquad i = 1, \dots, n.$$

Model obsahuje tři základní složky:

1. Pravděpodobnostní rozdělení vysvětlovaných proměnných.
2. Systematickou složku, která specifikuje vysvětlující proměnné.
3. Spojovací funkci (tzv. link) g, která popisuje funkční vztah mezi systematickou složkou a střední hodnotou vysvětlovaných proměnných.

Mezi zobecněné lineární modely patří například klasický lineární model, logistická regrese nebo log-lineární model. Jednou z velmi důležitých situací je logistická regrese. Ta je lineárním modelem pro logitovou transformaci binomických parametrů, viz 1.4.

1.3 Složky zobecněného lineárního modelu

Pod *náhodnou složkou* GLM rozumíme vysvětlovanou proměnnou Y. V modelu pracujeme s n realizacemi nezávislých náhodných veličin $Y_1, \ldots, Y_n$. Tato jednotlivá pozorování mají rozdělení z exponenciální rodiny. Tvar hustoty f tohoto rozdělení se dá zapsat jako:

$$f(y_i, \theta_i) = a(\theta_i) b(y_i) \exp[y_i Q(\theta_i)]. \tag{1.1}$$

Mezi důležité speciální případy patří např. normální, Poissonovo či binomické rozdělení.
Hodnota parametru θ_i může být obecně pro každé $i = 1, \ldots, n$ různá. To závisí na hodnotách matice vysvětlujících proměnných $\boldsymbol{X}$. Výraz $Q(\theta)$ se nazývá přirozený parametr.

Systematická složka se týká vektoru $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$ a vysvětlujících proměnných. Vztah mezi nimi se dá vyjádřit skrze lineární model. Nechť x_{ij} značí hodnotu vysvětlujících proměnných, $j = 1, \ldots, k$ pro každý jednotlivý případ i. Potom

$$\eta_i = \sum_{j=1}^{k} \beta_j x_{ij}.$$

Tato lineární kombinace vysvětlujících proměnných se nazývá lineární prediktor. Obvykle model zahrnuje také absolutní člen, tj. $x_{ij} = 1$ pro nějaké $j \in \{1, 2, \ldots, k\}$ pro všechna i (parametr u tohoto členu se pak často značí α nebo β_0). Uspořádáme-li tyto případy do matice, obdržíme tzv. regresní matici $\boldsymbol{X}$, která má n řádků a k sloupců a hodnost $\mathrm{h}(\boldsymbol{X}) = k > 0$, $n > k$.

Třetí složkou je tzv. *spojovací funkce* (*link*). Tato funkce udává vztah náhodné a systematické složky. Nechť $\mu_i = \mathrm{E}(Y_i)$, $i = 1, 2, \ldots, n$. Model spojuje μ_i s η_i a to vztahem $\eta_i = g(\mu_i)$. Funkce g je monotónní a diferencovatelná. Tedy g spojuje $\mathrm{E}(Y_i)$ s vysvětlujícími proměnnými. To můžeme vyjádřit pomocí vzorce:

$$g(\mu_i) = \sum_{j=1}^{k} \beta_j x_{ij} \qquad i = 1, 2, \ldots, n.$$

Spojovací funkce mohou mít obecně libovolný tvar, splněna musí být pouze podmínka spojitosti a diferencovatelnosti. Spojovací funkce $g(\mu) = \mu$, která se nazývá *identický link*, má tvar $\eta_i = \mu_i$. Tato spojovací funkce je užívaná

v klasické regresi s normálním rozdělení Y. Spojovací funkce, která transformuje střední hodnotu pomocí přirozeného parametru, se nazývá *kanonický link*. Dále uvádíme některé další spojovací funkce pro další modely:

- *identita*
$$\eta = \mu,$$
- *mocninné funkce*
$$\eta = \begin{cases} \mu^\lambda, & \text{pro } \lambda \neq 0, \mu > 0, \\ \log \mu, & \text{pro } \lambda = 0, \mu > 0, \end{cases}$$
- *komplementární log-log*
$$\eta = \log\{-\log(1-\mu)\}, \qquad 0 < \mu < 1,$$
- *probit*
$$\eta = \Phi^{-1}(\mu), \qquad 0 < \mu < 1,$$
kde $\Phi(.)$ je distribuční funkce normovaného normálního rozdělení,
- a pro binární data máme tzv. *logit*
$$\eta = \log\left(\frac{\mu}{1-\mu}\right).$$

1.4 Logistická regrese

Ve světě kolem nás se často setkáváme s vysvětlovanou proměnnou, která je 0-1 charakteru. Nejprve odvoďme obecný model pro situace, kdy vysvětlovaná proměnná nabývá pouze dvou možných hodnot.
Tedy platí $P(y = 1) = \pi$ a $P(y = 0) = 1 - \pi$. Střední hodnota je rovna $\mathrm{E}(Y) = \pi$. Hustota má pak tvar:

$$f(y, \pi) = \pi^y (1-\pi)^{1-y} = (1-\pi) \exp\left(y \log\left[\frac{\pi}{(1-\pi)}\right]\right). \tag{1.2}$$

To je hustota patřící do exponenciální rodiny ve tvaru (1.1), kde klademe

$$\theta = \pi, \tag{1.3}$$
$$a(\pi) = 1 - \pi, \tag{1.4}$$
$$b(y) = 1, \tag{1.5}$$
$$Q(\pi) = \log\left(\frac{\pi}{(1-\pi)}\right). \tag{1.6}$$

Podívejme se nyní blíže na statistický model binární vysvětlované proměnné. Výsledek každého takového pozorování je buď úspěch, nebo neúspěch. Binární data jsou pravděpodobně nejběžnější formou kategoriálních dat. Nejznámějším modelem pro binární data je logistická regrese.
Pro binární vysvětlovanou proměnnou Y a vektor kvantitativních vysvětlujících proměnných $\boldsymbol{x}$, nechť $\pi(\boldsymbol{x})$ značí pravděpodobnost, že náhodná veličina Y nabývá hodnoty 1 za podmínky $\boldsymbol{x}$, tedy:

$$\pi(\boldsymbol{x}) = \mathrm{P}(Y = 1|\boldsymbol{x})$$

Model logistické regrese je tvaru:

$$\mathrm{logit}[\pi(\boldsymbol{x})] = \log\left[\frac{\pi(\boldsymbol{x})}{(1-\pi(\boldsymbol{x}))}\right] = \alpha + \boldsymbol{x}'\boldsymbol{\beta}.$$

Tento vzorec můžeme alternativně přepsat jako přímé vyjádření pravděpodobnosti $\pi(\boldsymbol{x})$ jako:

$$\pi(\boldsymbol{x}) = \frac{\exp(\alpha + \boldsymbol{x}'\boldsymbol{\beta})}{1 + \exp(\alpha + \boldsymbol{x}'\boldsymbol{\beta})}.$$

Kapitola 2

Bradley-Terryho model

2.1 Definice Bradley-Terryho modelu

Bradley-Terryho model patří mezi logistické modely pro párová porovnávání (viz předchozí kapitola).
Uvažujme soubor o velikosti n různých pozorovaných prvků (hráčů, týmů apod.). Předpokládejme, že je možné porovnat každý prvek s jakýmkoli jiným. Zaveďme následující označení:

$$P(i \text{ zvítězí nad } j) = \pi_{ij}, \qquad i \neq j. \tag{2.1}$$

Je zřejmé, že pro n prvků existuje celkem $\binom{n}{2} = \frac{n(n-1)}{2}$ takových porovnávání. Dále předpokládejme, že při vzájemném porovnávání dvou různých prvků i a j mohou nastat pouze následující dvě možnosti:

1. i porazí j,
2. j porazí i.

To společně s (2.1) dává:

$$\pi_{ij} = 1 - \pi_{ji}. \tag{2.2}$$

To znamená, že nemůže nastat remíza.
Bradley-Terryho model předpokládá, že pro každý prvek i ze souboru o velikosti n existuje parametr β_i, který vyjadřuje jeho obecnou schopnost porazit ostatní prvky z našeho souboru (v příkladech tedy β_i vyjadřuje sílu fotbalového týmu, dovednost hráče tenisu nebo kvalitu vína). Dále předpokládá,

že při párovém porovnání se pravděpodobnost výhry i nad j dá vyjádřit následujícím způsobem:

$$\text{logit}(\pi_{ij}) = \beta_i - \beta_j. \tag{2.3}$$

Tedy z definice funkce logit a z jednoduchých úprav postupně dostaneme, viz [2]:

$$\log\left(\frac{\pi_{ij}}{\pi_{ji}}\right) = \beta_i - \beta_j, \tag{2.4}$$

$$\frac{\pi_{ij}}{1-\pi_{ij}} = \exp\left(\beta_i - \beta_j\right), \tag{2.5}$$

$$\pi_{ij} = \frac{\exp\beta_i}{\exp\beta_i + \exp\beta_j}. \tag{2.6}$$

Z poslední rovnosti si můžeme uvědomit dva jednoduché, ale užitečné důsledky:

- $P(i \text{ zvítězí nad } j) = \frac{1}{2} \iff \beta_i = \beta_j$,
- $P(i \text{ zvítězí nad } j) > \frac{1}{2} \iff \beta_i > \beta_j$.

Model v uvedené podobě nemá jednoznačně dané parametry $\beta_1, \beta_2, \ldots, \beta_n$. Proto se přidává podmínka restrikce na hodnoty jednoho či více parametrů β_i. V odborné literatuře a aplikacích Bradley-Terryho modelu se nejčastěji pokládá:

- $\beta_1 = 0$ nebo $\beta_n = 0$.
- $$\sum_{i=1}^{n} \exp\beta_i = 1. \tag{2.7}$$
- $\beta_i = 0$, kde $\beta_j > \beta_i$ pro $\forall j \neq i$. Dosáhneme tím toho, že parametry modelu jsou nezáporné.

Pokud jsme zvolili podmínku $\beta_i = 0$ pro nějaké $i \in \{1, 2, \ldots, n\}$, pak takový prvek i nazýváme jako *referenční*.

2.2 Modifikace Bradley-Terryho modelu

Do této chvíle nás zajímal pouze odhad pravděpodobnosti vítězství jednoho týmu nad druhým v jednom zápase. V této sekci popíšeme různé úpravy Bradley-Terryho modelu pro obecnější situace, jako jsou například odhady výsledků sérií zápasů. Dále zde zohledníme výhodu domácího prostředí nebo jiné specifické podmínky pro jednotlivá utkání. V neposlední řadě se budeme zabývat vlivem samotných hráčů na jejich tým a nakonec modifikujeme model do tvaru, kdy schopnosti hráčů modelujeme jejich osobními charakteristikami.

2.2.1 Bradley-Terryho model pro série zápasů

Předpokládejme situaci podobnou té z předchozí sekce. Máme n prvků, $\binom{n}{2}$ možností porovnání, kde n_{ij} značí počet opakování porovnání prvků i a j. Sílu každého prvku i určuje opět parametr β_i. Pro snadnější práci si zaveďme označení

$$\pi_i = \exp \beta_i, \qquad i = 1, 2, \ldots, n.$$

Dále použijeme podmínku (2.7) ve tvaru $\sum_{i=1}^{n} \pi_i = 1$ a dále $\pi_i \geq 0$, $i = 1, \ldots, n$. Jedná se tedy o pravděpodobnostní rozdělení. Pak platí:

$$P(i \text{ zvítězí nad } j) = \frac{\pi_i}{(\pi_i + \pi_j)}. \tag{2.8}$$

Vzorce pro maximální věrohodnost odhadující parametry π_i a vzorce pro test věrohodnostních poměrů jsou dány dle [3].
Pravděpodobnost pozorovaného výsledku v n_{ij} pozorování při porovnání prvků i a j je

$$\left(\frac{\pi_i}{\pi_i + \pi_j}\right)^{a_{ij}} \left(\frac{\pi_i}{\pi_i + \pi_j}\right)^{a_{ji}}, \tag{2.9}$$

kde definujeme

$$a_{ij} = 2n_{ij} - \sum_{k=1}^{n_{ij}} r_{ijk}, \tag{2.10}$$

$$a_{ij} + a_{ji} = n_{ij}, \tag{2.11}$$

kde $r_{ijk} = 1$, když i porazí j, a $r_{ijk} = 2$, když j porazí i, v k-tém opakování porovnání páru (i, j).
Poznamenejme, že a_{ij} je počet, kolikrát je upřednostněno i před j.

Násobením příslušných výrazů pro všechny opakování všech párů obdržíme výraz $L(\pi_i)$ pro známou věrohodnostní funkci. Tedy:

$$L(\pi_i) = \prod_i \pi_i^{a_i} \prod_{i<j} (\pi_i + \pi_j)^{-n_{ij}}, \tag{2.12}$$

kde $a_i = \sum_{j \neq i} a_{ij}$.
Rovnice (2.12) je analogická s rovnicí (2.8) danou Bradley-Terrym modelem, kterou získáme, když všechna n_{ij} jsou rovna n.

2.2.2 Bradley-Terryho model s výhodou domácího prostředí

V mnoha sportech se setkáváme se situací, kdy daný tým prokazuje na své domovské půdě lepší výsledky, než by dosahoval v utkáních na neutrálních hřištích. Tento aspekt zohledňuje verze Bradley-Terryho modelu, která přidává do modelu parametr modelující právě tuto výhodu domácího prostředí. Na danou situaci lze nahlížet i z druhé strany — pro hostující tým se jedná o znevýhodňující parametr.
Parametr určující výhodu domácího prostředí označme jako α.
Dále tedy pro $i \neq j$ nechť π^*_{ij} značí pravděpodobnost, že tým i porazí tým j za podmínky, že i je domácí tým. Upravená definice (2.13) tedy vypadá:

$$P(i \text{ zvítězí nad } j | i \text{ má domácí výhodu}) = \pi^*_{ij} \qquad i \neq j. \tag{2.13}$$

Základní model (2.4) rozšíříme následujícím způsobem, viz [1]:

$$\log\left(\frac{\pi^*_{ij}}{\pi^*_{ji}}\right) = \alpha + (\beta_i - \beta_j). \tag{2.14}$$

Je-li parametr α větší než nula, pak výhoda domácího prostředí existuje. Pokud je naopak α záporné, pak má domácí tým nevýhodu. Pokud platí $\alpha = 0$, potom místo konání zápasu nehraje roli.
Model je možné dále specifikovat. Neomezíme se pouze na výhodu domácího hřiště, ale vezmeme v potaz i další aspekty ovlivňující výsledek utkání. Toto jedno utkání je ovlivňováno více na sobě nezávislými faktory. Pro tyto faktory je nutné, aby se daly kvantifikovat a byly nezávislé na tom, jaké dva týmy se utkávají. Dále musí být zřejmé, že pro jeden z týmů je vždy tento

faktor výhodou.
Rozšířený model má tvar:

$$\log\left(\frac{\pi^*_{ij}}{\pi^*_{ji}}\right) = (\beta_i - \beta_j) + \sum_{t=1}^{M} \delta_t z_t. \tag{2.15}$$

Máme celkem M faktorů specifikující dané utkání, $z_1, z_2, \ldots, z_M$ jsou hodnoty nezávislých proměnných popisující tyto faktory, parametry $\delta_1, \delta_2, \ldots, \delta_M$ pak určují vliv faktorů na výsledek utkání. Například model (2.14) v tomto rozšířeném tvaru může vypadat následovně:

$$\log\left(\frac{\pi^*_{ij}}{\pi^*_{ji}}\right) = (\beta_i - \beta_j) + \delta_1 z_1,$$

kde $z_1 = 1$ pokud je tým i domácí, $z_1 = -1$ pokud tým i hostuje a parametr δ_1 je ekvivalentní s parametrem α v modelu (2.14).

Pro lepší představu si uveďme příklad: Volejbalový tým A hraje doma a vyhraje. Následující den hraje další utkání s jiným týmem a prohraje. Den nato hraje další utkání, které pro změnu vyhraje. Čtvrtý den po sobě odehraje čtvrté utkání opět s vítězným výsledkem. Pátý den se má utkat na domácí půdě s týmem B, který po posledním prohraném utkání týden odpočíval. A zde hrají roli právě ony další aspekty. Tým A sehrál čtyři utkání ve čtyřech dnech a na odpočinek má pouze den, zatímco tým B týden odpočíval. Únava hráčů týmu A je tedy poznatelně vyšší, to indikuje jistou výhodu pro tým B. Dále si uvědomme, že tým A vyhrál dva zápasy v řadě — to se obecně ve sportu považuje také za výhodu. Další výhodou pro tým A je domácí prostředí. Posledním aspektem je počet zraněných hráčů na obou stranách — dejme tomu, že díky dlouhé sérii zápasů má tým A zraněné čtyři hráče, zatímco tým B pouze jednoho.

Uvažujme model VOLEJBAL, který je definován následující rovnicí:

$$\log\left(\frac{\pi^*_{ij}}{\pi^*_{ji}}\right) = (\beta_i - \beta_j) + \delta_1 z_1 + \delta_2 z_2 + \delta_3 z_3 + \delta_4 z_4, \tag{2.16}$$

$z_1 = 1$ pokud je tým i domácí, $z_1 = -1$ pokud tým i hostuje,
$z_2 =$ (Počet dní odpočinku týmu i) − (Počet dní odpočinku týmu j),
$z_3 =$ (Počet zápasů od prohry týmu i) − (Počet zápasů od prohry týmu j),

$z_4 =$ (Počet zraněných hráčů týmu i) $-$ (Počet zraněných hráčů týmu j).

Dosadíme-li do tohoto modelu údaje vyplývající z předchozího textu, dostaneme model pro vítězství A nad B:

$$\log\left(\frac{\pi^*_{AB}}{\pi^*_{BA}}\right) = (\beta_A - \beta_B) + \delta_1 + (-6)\delta_2 + 2\delta_3 + (-3)\delta_4. \tag{2.17}$$

2.2.3 Bradley-Terryho model pro schopnosti hráčů

Představme si ligu v týmovém sportu, ve které jsou možné přestupy hráčů mezi týmy. Hráči tvoří množinu $H = \{1, 2, \ldots, n\}$.
V i-tém utkání ligy porovnáváme dvě disjunktní neprázdné podmnožiny této množiny T_i^+ a T_i^-. V řeči slov tedy T_i je množina hráčů, kteří se zúčastnili i-tého utkání. Je zřejmé, že hráč v utkání mohl nastoupit pouze za jeden z týmů. Máme tedy dvě disjunktní množiny T_i^+ a T_i^-. Každá z nich označuje soubor hráčů jednoho z týmů. To, že se utkání museli nějací hráči zúčastnit na obou stranách, netřeba dodávat, a tedy neprázdnost obou množin je zřejmá.
Ke každému utkání zaveďme dvě další proměnné Ψ a Ψ', pro něž platí:

$$\begin{array}{c}\text{Pokud } T_i^+ \text{ porazí } T_i^-, \text{ potom } \Psi_i = 1 \text{ a } \Psi'_i = 0,\\ \text{pokud } T_i^- \text{ porazí } T_i^+, \text{ potom } \Psi_i = 0 \text{ a } \Psi'_i = 1.\end{array}$$

Dále předpokládejme, že se odehraje P utkání.
Máme tedy $T_i \subset H$, $i = 1, 2, \ldots, P$:

$$T_i = T_i^+ \cup T_i^-, \qquad T_i^+ \cap T_i^- = 0, \qquad T_i^+ \neq 0, \qquad T_i^- \neq 0.$$

Předpokládáme, že všichni hráči jsou pro tým stejně důležití. Kvalita týmu je pak součtem kvalit všech jeho hráčů. Dostaneme následující model:

$$P(T_i^+ \text{ porazí } T_i^-) = \frac{\sum_{j \in T_i^+} \pi_j}{\sum_{j \in T_i} \pi_j}, \tag{2.18}$$

kde $\pi_j = \exp \beta_j$ značí kvalitu hráče j.

2.2.4 Bradley-Terryho model pro charakteristiky hráčů

V některých případech se setkáváme se situací, kdy máme k dispozici další vysvětlující proměnné charakterizující hráče. Pod tím si můžeme představit například výšku, váhu, věk či jiné měřitelné vlastnosti hráčů.
Pak lze parametry základního modelu (2.14) vyjádřit jako:

$$\beta_i = \sum_{r=1}^{p} \gamma_r x_{ir}, \tag{2.19}$$

kde je pro hráče i jeho dovednost vyjádřena jako součet vysvětlujících proměnných $x_{i1}, \ldots, x_{ip}$ s koeficienty $\gamma_1, \gamma_2, \ldots, \gamma_p$. V modelu (2.19) tedy odhadujeme právě koeficienty $\gamma_1, \gamma_2, \ldots, \gamma_p$.

2.2.5 Bradley-Terryho model zahrnující remízy

Stejně jako v odstavci 2.2.1 zavedeme označení

$$\pi_i = \exp \beta_i, \qquad i = 1, 2, \ldots, n.$$

Rao a Kupper, viz [10], navrhli pro data s remízami rozšíření základního modelu ve tvaru (2.6) s „threshold“ parametrem θ:

$$P(i \text{ zvítězí nad } j) = \frac{\pi_i}{\pi_i + \theta\pi_j}, \tag{2.20}$$

$$P(j \text{ zvítězí nad } i) = \frac{\pi_j}{\theta\pi_i + \pi_j}, \tag{2.21}$$

$$P(i \text{ remizuje s } j) = \frac{(\theta^2 - 1)\pi_i\pi_j}{(\theta\pi_i + \pi_j)(\pi_i + \theta\pi_j)}, \tag{2.22}$$

pro $i \neq j$, $i, j = 1, 2, \ldots, n$. Zřejmě platí:

$$P(i \text{ zvítězí nad } j) + P(j \text{ zvítězí nad } i) + P(i \text{ remizuje s } j) = 1. \tag{2.23}$$

Další alternativu modelu pracujícího s remízami představil Davidson, viz [4]. Při formulování modelu vyšel z axiomu známého pod pojmem „choice axiom“, viz [9]:

$$\frac{P(i \text{ zvítězí nad } j)}{P(j \text{ zvítězí nad } i)} = \frac{\pi_i}{\pi_j}, \tag{2.24}$$

pro všechna i, j splňující $P(i \text{ zvítězí nad } j) \neq 0, 1$. Davidson zavedl další podmínku pro pravděpodobnost remízy mezi i a j:

$$P(i \text{ remizuje s } j) = \nu\sqrt{P(i \text{ zvítězí nad } j)P(j \text{ zvítězí nad } i)}, \tag{2.25}$$

kde $\nu \geq 0$ je konstanta proporcionality, která nezávisí na i ani j.
Zkombinováním podmínek (2.24), (2.25) a (2.23) dostaneme model:

$$P(i \text{ zvítězí nad } j) = \frac{\pi_i}{\pi_i + \pi_j + \nu\sqrt{\pi_i\pi_j}}, \tag{2.26}$$

$$P(j \text{ zvítězí nad } i) = \frac{\pi_j}{\pi_i + \pi_j + \nu\sqrt{\pi_i\pi_j}}, \tag{2.27}$$

$$P(i \text{ remizuje s } j) = \frac{\nu\sqrt{\pi_i\pi_j}}{\pi_i + \pi_j + \nu\sqrt{\pi_i\pi_j}}. \tag{2.28}$$

Rozdíl mezi oběma modely je ten, že zatímco v rovnostech (2.20), (2.21) je oproti základnímu Bradley-Terry modelu přidán ve jmenovateli člen závisející na π_j (resp. π_i), u rovností (2.26), (2.27) je přidán prvek závisející na geometrickém průměru $\sqrt{\pi_i\pi_j}$.
Poznamenejme, že základní model (2.6) dostaneme z výše uvedených modelů, pokud položíme $\theta = 1$, resp. $\nu = 0$.

2.3 Odhady parametrů Bradley-Terryho modelu

Maximálně věrohodné odhady parametrů $\beta_1, \beta_2, \ldots, \beta_n$ základního modelu (2.14), případně odhady dalších parametrů v rozšířených modelech, nejdou obecně spočítat přímo. K jejich nalezení se proto používají iterativní numerické postupy, jejichž základy položil ve své práci před více než 80 lety Zermelo, viz [12].
V současné době jsou známy iterativní procedury pro odhad široké škály modifikací Bradley-Terryho modelu. Tyto postupy se však pro různé úpravy modelu mohou poměrně hodně lišit. Z tohoto důvodu uvedeme jen základní algoritmus pro odhad parametrů v základním modelu a nástin postupu pro odhad parametrů v modelu pro individuální schopnosti hráčů (2.18). Poskytnout ucelený přehled algoritmů pro různé modifikace Bradley-Terryho modelu by nebylo v rozsahu bakalářské práce možné, a ostatně to ani není cílem této práce.

Obecně postupujeme minimalizací negativní logaritmické funkce za určitých podmínek. Negativní logaritmickou funkcí rozumíme funkci, která vznikla ze známé věrohodnostní funkce. Věrohodnostní funkci zlogaritmujeme a vynásobíme -1. Označme opět

$$\pi_i = \exp \beta_i, \qquad i = 1, 2, \ldots, n.$$

Potom model (2.6) bude ve tvaru

$$\pi_{ij} = \frac{\pi_i}{\pi_i + \pi_j}. \tag{2.29}$$

Dále označme počet zápasů, kdy i porazilo j jako r_{ij}. Negativní logaritmická funkce je pak ve tvaru:

$$l(\boldsymbol{\pi}) = -\sum_{i<j} \left[r_{ij} \log \left(\frac{\pi_i}{\pi_i + \pi_j} \right) + r_{ji} \log \left(\frac{\pi_j}{\pi_i + \pi_j} \right) \right]. \tag{2.30}$$

Symbolem $\boldsymbol{\pi}$ značíme samozřejmě vektor parametrů $(\pi_1, \pi_2, \ldots, \pi_n)$.
Povšimněme si, že platí $l(\boldsymbol{\pi}) = l(\alpha\boldsymbol{\pi})$ pro všechna $\alpha > 0$. Proto budeme používat podmínku (2.7), tzn. $\sum_{i=1}^{n} \pi_i = 1$ a jakýkoli nenulový vektor $\boldsymbol{\pi}'$ můžeme tímto způsobem znormalizovat.

Odhady parametrů $\pi_1, \pi_2, \ldots, \pi_n$ získáme vyřešením optimalizační úlohy

$$\begin{array}{rl} \min_{\boldsymbol{\pi}} & l(\boldsymbol{\pi}) \\ \text{za podmínek} & \pi_i \geq 0, \qquad i = 1, 2, \ldots, n, \\ & \sum_{i=1}^{n} \pi_i = 1. \end{array} \tag{2.31}$$

Tuto úlohu lze vyřešit pomocí následujícího algoritmu:

1. Zvolme jakékoli hodnoty π_i^0, $i = 1, 2, \ldots, n$.
2. Opakujme pro $t = 0, 1, \ldots$
   (a) Nechť $s = (t \bmod n) + 1$.
   (b) Definujme

$$\boldsymbol{\pi}^{t+1} = \left(\pi_1^t, \pi_2^t, \ldots, \frac{\sum_{i \neq s} r_{si}}{\sum_{i \neq s} \frac{r_{si} + r_{is}}{\pi_s^t + \pi_i^t}}, \pi_{s+1}^t, \ldots, \pi_n^t \right)'.$$

(c) Normalizujme $\boldsymbol{\pi}^{t+1}$ ve smyslu, aby $\sum_{i=1}^{n} \pi_i^{t+1} = 1$.

3. Předchozí krok opakujeme do té doby, dokud není splněno

$$\partial l(\boldsymbol{\pi}^t)/\partial \pi_i < \epsilon, \qquad i = 1, 2, \ldots, n,$$

pro nějaké předem zvolené (malé) $\epsilon > 0$.
Pak algoritmus končí a $\hat{\boldsymbol{\pi}} = \boldsymbol{\pi}^t$ je odhadem vektoru parametrů v modelu (2.29).

Odhady dovedností jednotlivých hráčů jsou vzhledem k zavedenému značení rovny:

$$\hat{\beta}_i = \log \hat{\pi}_i, \qquad i = 1, 2, \ldots, n$$

a odhad pravděpodobnosti, že i porazí j je:

$$\hat{\pi_{ij}} = \frac{\hat{\pi}_i}{\hat{\pi}_i + \hat{\pi}_j}.$$

Podrobnější diskuzi k tomuto algoritmu lze nalézt například v [7]. Poznamenejme, že za splnění jistých (mírných) požadavků na koeficienty r_{ij} algoritmus konverguje k jedinému minimu úlohy (2.31) — více viz [8].

Nyní se budeme zabývat odvozením logaritmické funkce pro model s individuálními schopnostmi hráčů ze sekce 2.2.3. Připomeňme tedy model, se kterým budeme pracovat:

$$P(T_i^+ \text{ porazí } T_i^-) = \frac{\sum_{j \in T_i^+} \pi_j}{\sum_{j \in T_i} \pi_j},$$

kde T_i^+, T_i^- jsou disjunktní skupiny hráčů s dovednostmi π_j. Za předpokladu nezávislosti jednotlivých utkání pak celkovou dovednost týmů můžeme vyjádřit ve tvaru:

$$q_i = \sum_{j \in T_i = T_i^+ \cup T_i^+} \pi_j, \qquad q_i^+ = \sum_{j \in T_i^+} \pi_j, \qquad q_i^- = \sum_{j \in T_i^-} \pi_j. \tag{2.32}$$

Odhady parametrů modelu pak již dostaneme řešením optimalizační úlohy:

$$\begin{array}{rl} \min_{\boldsymbol{\pi}} & l(\boldsymbol{\pi}) = -\sum_{i=1}^{P} \left(\Psi_i \log \frac{q_i^+}{q_i} + \Psi_i' \log \frac{q_i^-}{q_i} \right) \\ \text{za podmínek} & \pi_i \geq 0, \qquad i = 1, 2, \ldots, n. \\ & \sum_{i=1}^{n} \pi_i = 1. \end{array} \tag{2.33}$$

Popis algoritmu řešící tuto úlohu lze nalézt opět v [7]. Metody pro nalezení maximálně věrohodných odhadů pro modely s remízami zavedených v sekci 2.2.5 zpracovává např. článek [4].
Takzvané MM algoritmy použitelné pro široké spektrum Bradley-Terry modelů společně s poznatky o konvergenci lze nalézt v práci D. Huntera [8].

Kapitola 3

Bradley-Terryho model v programu R

Tato kapitola práce si dává za úkol popsat práci s Bradley-Terryho modelem ve statistickém programu R, viz [14], tak, aby ji po přečtení zvládl i s Rkem dosud méně obeznámený uživatel. Výklad je průběžně ilustrován na práci s reálnými daty.

3.1 Instalace a příprava dat

Program R je volně šiřitelný statistický software, který lze zadarmo stáhnout z internetové adresy http://www.R-project.org. Podrobnější návod ke stažení a následné instalaci může čtenář nalézt například na stránkách RNDr. Arnošta Komárka, Ph.D. —
— http://www.karlin.mff.cuni.cz/∼komarek/Rko/Rmanual1.pdf.
Přídavný balík obsahující příkazy a funkce pro práci s Bradley-Terry modelem není součástí základní instalace programu R, proto je ho třeba nejprve nainstalovat. To se udělá příkazem:

```
> install.packages("BradleyTerry")
```

Poté (a při každém opětovném spuštění programu R) je třeba ještě zadat příkaz:

```
> library(BradleyTerry)
```

který umožňuje používat funkce z příslušného balíku.

3.1.1 Formát a příprava vstupních dat

Datový soubor pro Bradley-Terryho model v R musí obsahovat vždy nejméně dva sloupečky. V tom případě první řádek datového souboru bude typicky obsahovat tzv. hlavičku:

```
winner loser
```

Klíčové slovo `winner` nám říká, že v daném sloupci budeme uvádět jména vítězů daného zápasu, druhý sloupec bude pak obsahovat jména poražených (klíčové slovo `loser`).
Každý další řádek pak představuje výsledek jednoho zápasu — jak již bylo avizováno, na prvním místě uvedeme vítěze, na druhém pak jméno poraženého.
Pokud dva týmy sehrály více zápasů se stejným výsledkem, je možno přidat sloupec nadepsaný slovem `Freq`. Číslo v něm pak udává, kolik zápasů s daným výsledkem bylo sehráno.
Posledním volitelným sloupcem je sloupec obsahující údaje o tom, zda měl jeden z týmů nějakou výhodu (typicky domácího prostředí). Tento sloupec obsahuje hodnotu `1`, pokud byl vítěz daného zápasu (zápasů) pozitivně ovlivněn uvažovanou skutečností, hodnotu `-1`, jestliže byl vítěz ovlivněn negativně. Pokud nebyl zvýhodněn ani jeden hráč, pak bude hodnota v tomto sloupci rovna `0`.
Poznamenejme, že soubor s daty hlavičku obsahovat nemusí. Potom se automaticky bere první sloupec jako sloupec s vítězi a druhý jako sloupec poražených.
Příklad řekne víc než tisíc slov — představme si, že fotbalové kluby Sparta a Slavia spolu odehrají tři zápasy: první z nich vyhraje na domácí půdě Sparta, druhý z nich vyhraje na domácí půdě Slavia a ve třetím zápase na stadionu Sparty opět vyhraje Slavia. Následující tabulka ukazuje několik způsobů, jak tyto výsledky zapsat takovým způsobem, abychom je mohli použít v Bradley-Terryho modelu v prostředí R:

Sparta	Slavia
Slavia	Sparta
Slavia	Sparta

winner	loser	Freq
Sparta	Slavia	1
Slavia	Sparta	2

winner	loser
Sparta	Slavia
Slavia	Sparta
Slavia	Sparta

winner	loser	Freq	dom.vyhoda
Sparta	Slavia	1	1
Slavia	Sparta	1	1
Slavia	Sparta	1	-1

Pokud máme data připravena v příslušném formátu, do prostředí R je načteme příkazem:

```
> data = read.table(cesta souboru)
```

pokud soubor obsahuje hlavičku, potom příkazem:

```
> data = read.table(cesta souboru, header=TRUE)
```

Pokud tedy budeme mít umístěn soubor s daty data.txt přímo na disku C, načteme ho příkazem:

```
> data = read.table(''C:\data.txt'')
```

3.2 Funkce `BTm()`

Funkce a procedury, které obsahuje balík `BradleyTerry`, budeme ilustrovat na reálných datech - výsledcích zápasů mezi osmi nejlepšími tenisty světa podle žebříčku ATP k 1.dubnu 2010 za poslední tři roky. Zápasové bilance tenistů mezi sebou shrnuje následující tabulka, údaje v ní jsou získány ze stránek organizace ATP, [15]:

Federer	7-5	4-5	4-7	6-2	5-2	8-0	7-1
	Djokovic	3-3	7-13	3-0	3-2	5-1	2-4
		Murray	3-7	5-1	5-2	0-1	4-2
			Nadal	4-3	3-5	2-2	4-2
				Del Potro	1-3	2-1	3-0
					Davydenko	2-6	1-2
						Soderling	2-1
							Roddick

Tato data jsme upravili a načetli (viz 3.1.1) do prostředí R do proměnné `tenis`.

```
> tenis
      winner     loser Freq
1    Federer  Djokovic    7
2    Federer    Murray    4
3    Federer     Nadal    4
4    Federer  DelPotro    6
5    Federer Davydenko    5
...
55   Roddick Soderling    1
56   Roddick   Federer    1
```

Základní Bradley-Terryho model (2.3) odhadneme pomocí metody maximální věrohodnosti (viz sekce 2.3) pomocí funkce `BTm` a výsledky uložíme jako proměnnou `model1`:

```
> model1=BTm(formula= tenis~..)
> model1

Call:  BTm(formula = tenis ~ ..)

Coefficients:
 ..DelPotro   ..Djokovic    ..Federer     ..Murray      ..Nadal    ..Roddick
   -0.14305      0.41265      0.88631      0.59092      0.81264     -0.27550
..Soderling
   -0.08194

Degrees of Freedom: 28 Total (i.e. Null);  21 Residual
Null Deviance:      49.47
Residual Deviance: 31.82        AIC: 99.28
```

Kvůli identifikaci (viz 2.7) je jeden parametr modelu položen rovný nule, $\beta_i = 0$ pro nějaké i. V případě, že chceme zvolit, který hráč bude referenční, nastavíme parametr `refcat`:

```
> model1=BTm(formula=tenis~.., refcat="Roddick")
> model1

Call:  BTm(formula = tenis ~ .., refcat = "Roddick")

Coefficients:
..Davydenko   ..DelPotro   ..Djokovic    ..Federer     ..Murray      ..Nadal
     0.2755       0.1324       0.6881       1.1618       0.8664       1.0881
..Soderling
     0.1936

Degrees of Freedom: 28 Total (i.e. Null);  21 Residual
Null Deviance:      49.47
Residual Deviance: 31.82        AIC: 99.28
```

Roddick je podle Bradley-Terryho modelu nejhorším hráčem, proto jsou odhady všech parametrů v novém modelu nezáporné. Poznamenejme, že pokud je parametr `refcat` nespecifikován, vezme R jako referenční objekt (hráče) ten, který je ze všech první v lexikografickém uspořádání - stejně tak jako si seřadí i ostatní hráče. Pokud bychom chtěli dále pracovat s odhadnutým parametrem dovednosti hráče Davyděnko (ten je podle abecedy první ze všech hráčů), získáme ho pomocí příkazu:

```
> model1$coefficients[1]
..Davydenko
  0.2754978
```

Odhad pravděpodobnosti, že i-tý hráč porazí j-tého, spočítáme podle vzorce (2.6). Tedy například pravděpodobnost, že Federer (4.) porazí Murrayho (5.), spočítáme v R následovně:

```
> exp(model1$coefficients[[4]])/(exp(model1$coefficients[[5]])+exp(model1$coefficients[[4]]))
[1] 0.573316
```

To znamená, že odhad této pravděpodobnosti v Bradley-Terryho modelu je přibližně 57%. Stejným způsobem je možné spočítat odhady všech pravděpodobností:

Federer	0.61	0.57	0.51	0.73	0.70	0.72	0.76
	Djokovic	0.45	0.40	0.63	0.60	0.62	0.66
		Murray	0.44	0.67	0.64	0.66	0.70
			Nadal	0.72	0.69	0.70	0.74
				Del Potro	0.46	0.48	0.53
					Davydenko	0.52	0.56
						Soderling	0.54
							Roddick

Pro model vytvořený pomocí funkce `BTm` jsou použitelné standardní procedury třídy `glm`, proto si některé základní charakteristiky našeho modelu, jako odhady parametrů, jejich směrodatné odchylky a p-hodnoty, můžeme prohlédnout zavoláním procedury `summary`:

```
> summary(model1)

Call:
BTm(formula = tenis ~ .., refcat = "Roddick")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2695  -0.6783   0.1417   0.8764   1.6591

Coefficients:
```

```
            Estimate Std. Error z value Pr(>|z|)
..Davydenko   0.2755     0.4734   0.582  0.56063
..DelPotro    0.1324     0.4983   0.266  0.79038
..Djokovic    0.6881     0.4363   1.577  0.11473
..Federer     1.1618     0.4313   2.694  0.00707 **
..Murray      0.8664     0.4540   1.908  0.05635 .
..Nadal       1.0881     0.4371   2.489  0.01279 *
..Soderling   0.1936     0.4999   0.387  0.69863
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 49.472  on 28  degrees of freedom
Residual deviance: 31.821  on 21  degrees of freedom
AIC: 99.285

Number of Fisher Scoring iterations: 4
```

3.2.1 Parametr `order.effect` (výhoda domácího prostředí)

Rozšíření Bradleyho-Terryho modelu používáme, když předpokládáme, že výsledek sportovního utkání ovlivňuje místo, kde se utkání hraje, nebo pořadí, v jakém someliér ochutnává víno (proto se procedura nazývá `order.effect`) - více viz 2.2.2). Parametr `order.effect` ve volání funkce `BTm` musí být rovný vektoru obsahujícímu hodnoty z množiny {1,-1,0}, který má stejnou délku, jako je počet řádků seznamu výsledků utkání.
Výchozí datový soubor z výsledky utkání jsme proto museli tímto způsobem upravit. Federer porazil Djokoviče sedmkrát, z toho šestkrát na neutrální půdě a jednou v rodném Švýcarsku. Proto první řádek s výsledky utkání v původním souboru:

```
     winner     loser Freq
1   Federer  Djokovic    7
...
```

se rozpadne na dva:

```
     winner     loser Freq domvyh
1   Federer  Djokovic    6  0
2   Federer  Djokovic    1  1
...
```

Obdobně jsme upravili i ostatní záznamy. Bradley-Terryho model zohledňující výhodu domácího prostředí nyní odhadneme následovně:

```
> modelDV=BTm(tenis~.., order.effect=tenis$domvyh)
> summary(modelDV)

Call:
BTm(formula = tenis ~ .., order.effect = tenis$domvyh)

Deviance Residuals:
     Min       1Q   Median       3Q      Max
-1.31729  -0.30940  -0.02369   0.64034   1.40350

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
..DelPotro   0.01298    0.68713   0.019   0.9849
..Djokovic   0.03350    0.63577   0.053   0.9580
..Federer    0.24620    0.62765   0.392   0.6949
..Murray    -0.18064    0.64544  -0.280   0.7796
..Nadal      0.28488    0.62493   0.456   0.6485
..Rodick    -0.51735    0.63048  -0.821   0.4119
..Soderling  0.13808    0.67885   0.203   0.8388
.order       0.74982    0.42195   1.777   0.0756 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 30.498  on 51  degrees of freedom
Residual deviance: 24.855  on 43  degrees of freedom
AIC: 81.058

Number of Fisher Scoring iterations: 3
```

Koeficient vlivu domácího prostředí (`.order`) je 0.749. Jak z výstupu vidíme, v porovnání s koeficienty u jednotlivých tenistů je to vysoká hodnota a můžeme tedy prohlásit, že výhoda domácího hřiště hraje v tenise velkou roli. Přesněji, šance se pro tenistu přechodem z neutrálního na domácí kurt zvedne asi 2.11-krát ($\exp 0.749 = 2.11488$).

3.2.2 Použití vysvětlujících proměnných

V některých případech můžeme každému hráči přiřadit několik vysvětlujících proměnných. Parametry β_i jsou pak ve tvaru:

$$\beta_i = \sum_{r=1}^{p} \gamma_r x_{ir}, \tag{3.1}$$

více viz (2.2.4). Funkce `BTm` v takovém případě odhaduje parametry $\gamma_1, \ldots, \gamma_p$. Vše si opět ukážeme na pokračování našeho příkladu. Musíme mít připravená data, ve kterých jsou uvedeny vysvětlující proměnné u každého z hráčů, který se vyskytuje v seznamu výsledků utkání. V našem případě budeme

uvažovat tři vysvětlující proměnné - dvě kvantitativní proměnné věk a výška hráče a jednu binární proměnnou - to, zda hráč hraje levou, či pravou rukou. Tyto údaje shrnuje následující tabulka:

```
> hraci
          Vyska Ruka Vek
Federer     185    R  28
Djokovic    188    R  22
Murray      190    R  22
Nadal       185    L  23
DelPotro    198    R  21
Davydenko   175    R  28
Soderling   193    R  25
Roddick     188    R  27
```

Bradley-Terryho model s prediktory `Vyska`, `Ruka`, `Vek` odhadneme v programu R následovně:

```
> model2=BTm(tenis~Vyska+Vek+Ruka, data=hraci)
> summary(model2)

Call:
BTm(formula = tenis ~ Vyska + Vek + Ruka, data = hraci)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2156  -0.7820   0.4690   0.9116   1.9537

Coefficients:
        Estimate Std. Error z value Pr(>|z|)
Vyska -0.0116496  0.0240051  -0.485    0.627
Vek    0.0009776  0.0521714   0.019    0.985
RukaR -0.4416487  0.2800438  -1.577    0.115

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 49.472  on 28  degrees of freedom
Residual deviance: 45.475  on 25  degrees of freedom
AIC: 104.94

Number of Fisher Scoring iterations: 4
```

Tedy hráči hrající levou rukou mají nad pravorukými $\exp 0.44 = 1.55$krát větší šanci na vítězství a šance na vítězství klesá s narůstající výškou.

V případě, že u nějakého hráče některá data chybí, není vždy situace ztracená a zkombinováním rovnice (3.1) a (2.3) se v některých situacích můžeme obejít i bez chybějících dat. Pokud například u k-tého hráče chybí

jedna nebo více vysvětlujících proměnných $x_{k1}, \ldots, x_{kp}$, potom logit model můžeme vyjádřit jako:

$$\text{logit}(P(\text{k porazí j})) = \beta_k - \sum_{r=1}^{p} \gamma_r x_{jr}, \tag{3.2}$$

pro $j \neq k$.

3.2.3 Ostatní parametry

U funkce `BTm()` můžeme, podobně jako například u funkce `glm()`, použít parametr `subset` pro nastavení podmnožiny dat, na které budeme odhad provádět, či `contrasts` nebo `offset` pro nastavení kontrastů, resp. ofsetu v modelu.
Užitečný může být speciální binární parametr `br`, který je defaultně nastaven na hodnotu `FALSE`. Pokud máme k dispozici balíček *brlr*, nastavením `br=TRUE` dostaneme maximálně věrohodný odhad s redukovaným vychýlením, viz [5].

3.3 Další funkce balíku BradleyTerry

Další funkce v balíku BradleyTerry jsou:

`BTabilities()` - Vypíše odhady dovedností jednotlivých hráčů β_j a směrodatné chyby těchto odhadů:

```
> BTabilities(model1)
             ability      s.e.
Davydenko  0.00000000 0.0000000
DelPotro  -0.14305005 0.4627340
Djokovic   0.41264823 0.4035474
Federer    0.88630879 0.3971484
Murray     0.59091554 0.4179322
Nadal      0.81264416 0.3923629
Roddick   -0.27549777 0.4734367
Soderling -0.08193743 0.4395235
```

`BTresiduals()` - tato funkce spočítá rezidua, která odhadují chybu v lineárním prediktoru $\sum \gamma_r x_{ir}$ (hodí se tedy při modelování pomocí vysvětlujících

proměnných, viz odstavec 3.2.2). Při použití této funkce na našem příkladu s tenisty dostaneme:

```
BTresiduals(model2)
    Davydenko      DelPotro      Djokovic       Federer        Murray
-4.713399e-01 -3.647117e-01  1.029789e-01  6.437348e-01  3.713398e-01
        Nadal       Roddick     Soderling
 5.137808e-11 -6.434489e-01 -2.789447e-01
attr(,"weights")
Davydenko  DelPotro  Djokovic   Federer    Murray     Nadal   Roddick Soderling
10.376344  8.313249 14.212509 15.595695 11.079143 15.366152  8.423927  8.152574
```

Tato rezidua jsou získána výpočtem z tzv. „working“ reziduí z odhadu modelu pomocí `glm`. Parametr `weights` vyjadřuje informaci vyjádřenou v jednotlivých reziduích - jsou nepřímo úměrné rozptylu - a vážená regrese s těmito vahami na jakoukoli vysvětlující proměnnou bude vykazovat odhady parametrů blížící se nule.
Z těchto reziduí můžeme například okamžitě vyčíst, že `model2` podhodnocuje dovednost Roddicka a naopak nadhodnocuje umění Federera. Někdy může být užitečné i grafické zobrazení reziduí (resp. koeficientů `weights`) v závislosti na některé z vysvětlujících proměnných.

`add1.BTm()`, `drop1.BTm()` - balíček `BradleyTerry` umožňuje užití i těchto funkcí, které nám mohou pomoci při vyřazení či zařazení vysvětlujících proměnných do modelu. Tím pádem lze vhodný podsoubor vysvětlujících proměnných najít například pomocí funkce `step()`. Návod k jejímu použití je možno najít v [13].
Kompletní popis balíku `BradleyTerry` lze nalézt v [6].

3.4 Nedostatky balíku BradleyTerry

Mezi hlavní nedostatky balíku `BradleyTerry` můžeme považovat:

- Skutečnost, že nejsou implementovány modely, které počítají i s remízami. (To se dá řešit způsobem, že zápasy, které skončí remízou, jednoduše vypustíme, alternativně tak, že každému ze zúčastněných týmů připíšeme půl vítězství - viz [11].)

- Až na výhodu domácího prostředí (viz odstavec 3.2.1) není možné pracovat s proměnnými, které popisují specifické podmínky jednotlivých utkání (např. počasí).

- Z pohledu autora práce i to, že funkce `BTm` požaduje data v poměrně striktní podobě. Příprava dat z běžně dostupných zdrojů proto není triviální záležitostí a může vyžadovat větší množství času.

Některé z popsaných nedostatků ovšem řeší balíček `BradleyTerry2` - viz sekce 3.5.

3.5 Balíček BradleyTerry2

V době psaní této práce Heather Turner a David Firth z university ve Warwicku, UK, pracovali na přídavném balíku do R - BradleyTerry2, viz [11].

Nejvýraznějším vylepšením tohoto balíku oproti balíčku BradleyTerry, je možnost počítat model ve tvaru:

$$\beta_i = \sum_{r=1}^{p} \gamma_r x_{ir} + U_i, \tag{3.3}$$

kde $U_1, \ldots, U_n$ jsou nezávislé stejně rozdělené náhodné veličiny, $U_i \sim N(0, \sigma_U)$ pro všechna $i = 1, \ldots, p$.
Důležité je také rozšíření funkce `BTm()`, díky kterému můžeme v modelu uvažovat více vysvětlujících proměnných spojených s podmínkami zápasu (contest-specific predictors) - v balíčku BradleyTerry byla použitelná pouze výhoda domácího prostředí - viz `order.effect`, 3.2.1. Nyní tak při modelování výsledků sportovních zápasů můžeme využít i další údaje, jako je například délka odpočinku před zápasem.
Dále balíček umožňuje lepší práci s daty (např. funkce `countsToBinomial()` převede křížovou tabulku výsledků na požadovaný formát dat), dalšími způsoby rozšiřuje funkci `BTm()` a nabízí několik dalších nových funkcí. Stejně jako balíček BradleyTerry si nedokáže poradit s daty, které obsahují remízy.

K 2.dubnu 2010 byl balík BradleyTerry2 k dispozici ve verzi 0.9-2.

Kapitola 4

Analýza dat ve volejbalové lize

4.1 Úvod

Pro analýzu jsme si vybrali nejvyšší volejbalovou soutěž v České republice. Volejbal je míčový sport pro dva týmy. Na každé straně proti sobě stojí šest hráčů. Cílem hry je dostat míč přes síť na soupeřovo pole tak, aby se dotkl země. Hraje se vždy na tři vítězné sety. Hraje se beze ztrát, tedy když dané mužstvo úspěšně složí míč do hřiště soupeře, popřípadě soupeř odehraje míč do zámezí bez přičinění druhého mužstva, získává mužstvo bod. V setu se vždy hraje do 25 bodů, přičemž se musí zvítězit o dva body. V případném pátém setu se hraje pouze do 15 bodů, nicméně pravidlo zvítězit o dva body zůstává zachováno. Nejvyšší volejbalovou soutěž hrálo v letech 2005-2008 vždy deset týmů. Při hracím systému každý s každým odehrál každý tým 27 soutěžních utkaní v základní části. Pak následovalo vždy play off hrané na tři vítězná utkání. My se ovšem zabýváme pouze základní částí.

4.2 Popis dat

V následující analýze budeme používat výsledky a údaje o utkáních v české volejbalové extralize ze tří sezon: 2005/06, 2006/07 a 2007/08. Data jsme získali ze stránek českého volejbalového svazu [16] a sportovní redakce serveru iDnes [17]. V každé sezoně se ligy zúčastnilo deset týmů. Hrací systém byl následující: každý s každým odehrál tři zápasy, z toho jeden na domácí půdě, druhý na hřišti soupeře a místo dějiště třetího vzájemného utkání bylo určeno na základě regulí losu. S každé sezony máme tedy 27 zápasů

Ostrava	7-2	4-5	1-8	7-2	6-0	3-0	7-2	9-0	9-0	6-0
	Liberec	5-4	6-3	8-1	5-4	3-0	8-1	8-1	7-2	5-1
		České	4-5	2-7	7-2	3-0	7-2	8-1	8-1	6-0
		Budějovice	Kladno	5-4	5-4	5-1	7-2	6-3	6-3	5-1
				Opava	6-3	1-2	6-3	5-4	7-2	6-0
					Ústí	1-2	3-6	5-4	3-6	5-1
					n. L.	ČZU	2-1	2-1	2-1	0-0
						Praha	Brno	4-5	4-5	3-3
								Zlín	5-4	3-3
									Příbram	2-4
										Benátky

Tabulka 4.1: *Tabulka vzájemných zápasů. Většina týmů má devět vzájemných zápasů, pouze Benátky nad Jizerou se s každým týmem kromě ČZU Praha utkali šestkrát, tým z Prahy pak s každým kromě Benátek třikrát.*

pro každý tým. V jedné sezoně se tedy odehrálo celkem 135 zápasů, za tři sezony tak máme údaje o 405 zápasech.
V sezoně 2005/06 a 2006/07 hrály extraligu týmy Ostravy, Liberce, Kladna, Opavy, Č.Budějovic, Ústí nad Labem, Zlína, Brna, Příbrami a Benátek nad Jizerou. Poslední jmenovaní sestoupili, a tak je na následující sezonu nahradil tým ČZU Praha. Devět týmů, které odehrály všechny tři sezony, mají každý s každým devět vzájemných zápasů, s Benátkami 6 zápasů a ČZU Praha 3 vzájemné zápasy. Benátky a ČZU spolu ve sledovaném období nehrály ani jeden extraligový zápas.
V tabulce 4.1 jsou shrnuty vzájemné zápasy všech týmů. V tabulce 4.2 je potom procento vítězných utkání v extralize pro všechny týmy.

Kromě výsledků vzájemných utkání jsme tedy získali i různé další proměnné, které charakterizují jednotlivé týmy. Vybrali jsme následující veličiny:

- `Vek` - průměrný věk hráčů v týmu.
- `Vaha` - průměrná váha hráčů v týmu.
- `Vyska` - průměrná výška hráčů v týmu.
- `Divaci` - průměrný počet diváků na domácích zápasech týmu.

U veličin týkajících se hráčů se do průměru počítá vždy jen 12 hráčů ze soupisky, kteří mají nejvíce odehraných zápasů.
Údaje ke všem týmům hrajícím českou nejvyšší volejbalovou soutěž jsou shrnuté v tabulce 4.3. Údaje jsou ze sezóny 2007/08. Minima a maxima jednotlivých veličin jsou v ní vyznačena tučně. Průměrný věk v soutěži je 25.89 let. S přehledem nejstarší sestavu měl tým z Benátek nad Jizerou (průměrný věk

	2005/06	*Z*	*V*	*P*	*sety*	*B*
1.	Ostrava	27	22	5	73:20	49
2.	Liberec	27	19	8	64:43	46
3.	Kladno	27	19	8	65:45	46
4.	Opava	27	17	10	56:45	44
5.	Č.Budějovice	27	14	13	58:56	41
6.	Ústí n. L.	27	11	16	44:61	38
7.	Zlín	27	10	17	42:59	37
8.	Brno	27	9	18	42:62	36
9.	Příbram	27	9	18	44:63	36
10.	Benátky n. J.	27	5	22	36:70	32

	2006/07	*Z*	*V*	*P*	*sety*	*B*
1.	České Budějovice	27	21	6	67:31	48
2.	Ostrava	27	20	7	67:37	47
3.	Liberec	27	19	8	68:43	46
4.	Opava	27	16	11	60:43	43
5.	Kladno	27	13	14	58:52	40
6.	Ústí n. L.	27	11	16	40:64	38
7.	Brno	27	10	17	47:59	37
8.	Zlín	27	10	17	43:59	37
9.	Příbram	27	9	18	46:65	36
10.	Benátky n. J.	27	6	21	28:71	33

	2007/08	*Z*	*V*	*P*	*sety*	*B*
1.	Liberec	27	19	8	68:40	46
2.	Kladno	27	19	8	65:39	46
3.	Ostrava	27	18	9	64:45	45
4.	Č. Budějovice	27	17	10	64:39	44
5.	Opava	27	13	14	54:59	40
6.	Brno	27	12	15	48:54	39
7.	Ústí n. L.	27	11	16	48:63	38
8.	ČZU Praha	27	10	17	45:64	37
9.	Zlín	27	8	19	45:64	35
10.	Příbram	27	8	19	42:64	35

tým	*procento vítězství*
Ostrava	68,97%
Liberec	65,52%
Č.Budějovice	59,77%
Kladno	58,62%
Opava	52,87%
Ústí nad Labem	37,93%
ČZU Praha	37,04%
Brno	35,63%
Zlín	32,18%
Příbram	29,89%
Benátky nad Jizerou	20,37%

Tabulka 4.2: *Výsledné tabulky české nejvyšší volejbalové soutěže pro sezony 2005/06, 2006/07 a 2007/08. Vpravo dole jsou uvedena procenta vítězných zápasů pro všechny týmy, které se v těchto třech sezonách extraligy zúčastnily.*

29.56 let), za benjamínky soutěže můžeme označit mužstvo Kladna (23.55). Jak tedy vidíme, v tomto parametru mezi mužstvy panují poměrně znatelné rozdíly. Nejvyšším družstvem disponovaly Liberec a Opava (jejichž hráči byli průměrně vysocí shodně 194.92 cm), naopak trpaslíky v lize byli hráči Příbrami (193.73 cm). Všechna mužstva jsou natěsnána v jednom centimetru výšky (průměrná výška v soutěži je pak 194.5 cm). O něco větší rozdíly mezi mužstvy můžeme zaznamenat ve váze - ta se pohybuje mezi 86.59 až 90.83 kg. Nejtěžším mužstvem bylo mužstvo Ostravy, nejlehčími byli hráči Brna. Samostatnou kapitolou je návštěvnost diváků. Nejvíce lidí chodilo do haly v Českých Budějovicích, průměrně to bylo 697 diváků na zápas. Naopak nejméně lákalo mužstvo Kladna, kdy do kladenské haly na zápas domácích našlo cestu průměrně pouze 338 diváků. V celé extralize pak činila průměrná divácká návštěvnost 479.1 diváka.
Zkoumat budeme tři modely, pro každý použijeme jiný soubor s daty. V prvním se budeme zabývat pouze tím, který z dvojice týmů vyhrál, bez ohledu na místo konání zápasu. Druhý model bude zohledňovat výhodu (nevýhodu) domácího prostředí. Tyto dva modely budou brát v úvahu všechny tři sezony. Konečně poslední, třetí model, se pokusí studovat vliv i jiných charakteristik jednotlivých týmů, v našem případě váhy, výšky, věku hráčů a návštěvnosti na domácích utkáních. Tento model bude na rozdíl od prvních dvou pracovat pouze se sezonou 2007/08.

Tým	Divaci	Vyska [cm]	Vaha [kg]	Vek
Benátky nad Jizerou	386	194.25	88.69	**29.56**
Brno	446	194.29	**86.59**	23.76
CB	**697**	194.67	88.5	28.00
Ostrava	500	194.25	**90.83**	29.00
Kladno	**338**	194.36	89.92	**23.55**
Liberec	551	**194.92**	90.08	25.45
Opava	421	**194.92**	87.75	24.55
CZU	427	194.66	88.8	24.72
Příbram	573	**193.73**	89.73	25.75
Ústí	481	194.67	88.8	24.83
Zlín	450	194.65	89.23	25.64

Tabulka 4.3: *V této tabulce jsou údaje o průměrné výšce, váze a věku hráčů v jednotlivých týmech a průměrné domácí návštěvy. Minima a maxima údajů jsou vyznačena tučně.*

4.3 Formulace modelů a odhad parametrů

V prvním modelu uvažujeme pouze vítězství či porážku daného týmu v utkání. V modelu odhadneme koeficienty $\beta_1, \ldots, \beta_{11}$ vyjadřující síly jednotlivých týmů v základním Bradley-Terryho modelu:

$$\text{logit}(\pi_{ij}) = \beta_i - \beta_j. \quad \text{MODEL 1}$$

V tabulce 4.4 jsou odhadnuté koeficienty. Z tabulky vidíme, že největší sílu podle Bradley-Terryho modelu má tým Ostravy, následován týmem Liberce a na třetím a čtvrtém místě jsou s téměř stejnou odhadnutou silou týmy Kladna a Budějovic. Na chvostu se potom umístilo družstvo Benátek nad-Jizerou a týmy z Příbrami a Zlína. Tabulka je seřazena podle odhadnuté síly týmů, pokud ji porovnáme s tabulkou procentuální úspěšnosti vítězství 4.2, zjistíme, že pořadí se až na výměnu Kladna s Českými Budějovicemi shoduje.
Ve druhém sloupci tabulky jsou spočítané směrodatné odchylky příslušných odhadů $\beta_1, \ldots, \beta_{11}$. V posledním sloupci můžeme nalézt p-hodnoty určující

MODEL 1	odhad	sm. odchylka	p-hodnota	
Ostrava	1.58221	0.34822	5.53e-06	***
Liberec	1.38516	0.33452	3.46e-05	***
Kladno	1.09348	0.32913	0.000893	***
CB	1.09289	0.32823	0.000870	***
Opava	0.81474	0.32072	0.011075	*
Ústí	0.15199	0.32027	0.635086	
CZU	0.05905	0.46789	0.899575	
Brno	0	0	0	
Zlín	-0.10008	0.32159	0.755641	
Příbram	-0.21933	0.32645	0.501657	
Benátky nad Jizerou	-0.79035	0.40602	0.051584	.

Tabulka 4.4: *Koeficienty $\beta_1, \ldots, \beta_{11}$ MODELu 1 odhadnuté v programu R. V prvním sloupci je uveden tým příslušný danému koeficientu.*

signifikanci proměnných v modelu. Obvykle se za hraniční hodnotu považuje 0.05 nebo 0.10. Pokud p-hodnota přesáhne tuto hladinu, pak příslušnou proměnnou můžeme prohlásit za nevypovídající a případně ji vyřadit z modelu. Přesněji řečeno to znamená, že příslušný parametr není statisticky odlišný od 0, tedy síla daného týmu se významně neliší od síly referenčního týmu. Z tabulky vidíme, že u parametrů některých týmů p-hodnota avizovanou hranici výrazně překročila. To přičítáme vyrovnanosti týmů, kterých se to týká a blízkostí odhadnutých koeficientů 0. U ČZU Praha tomu může být i z důvodu menšího počtu pozorovaných výsledků. Kvůli identifikaci modelu je referenčnímu prvku (kterým je Brno) přiřazena hodnota odhadu 0, proto je nulová i směrodatná odchylka a p-hodnota.

Dále jsme použili rozšířený Bradley-Terryho model zahrnující možnost výhody domácího prostředí:

$$\text{logit}(\pi_{ij}) = \alpha + (\beta_i - \beta_j). \quad \text{MODEL 2}$$

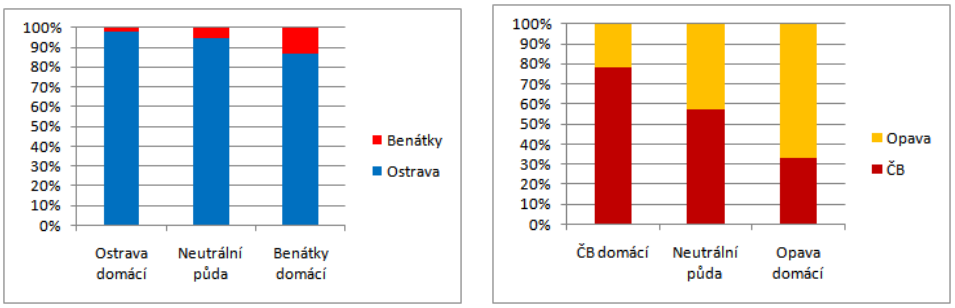
Odhadnuté koeficienty tohoto modelu jsou vypsány v tabulce 4.5. Oproti MODELu 1 je zde několik rozdílů. Nejvýznamnější je přítomnost parametru pro výhodu domácího prostředí, jehož odhadnutá hodnota, směrodatná odchylka a p-hodnota jsou v posledním řádku tabulky. Na první pohled je zřejmé, že by bylo chybou efekt domácího prostředí neuvažovat. To vyvozu-

jeme z faktu, že p-hodnota u tohoto odhadu je velmi blízká nule, což naznačuje statistickou významnost. Další významnou skutečností je vysoká hodnota odhadnutého vlivu domácího prostředí na výsledek zápasu. P-hodnoty u některých odhadů jsou opět výrazně vyšší než 0.05 - to přičítáme vyrovnanosti týmů, kterých se to týká a blízkostí odhadnutých koeficientů 0. U ČZU Praha tomu může být i z důvodu menšího počtu pozorovaných výsledků. V porovnání s tabulkou skutečné procentuální úspěšnosti 4.2 zůstává pořadí týmů stejné, pouze se prohodilo pořadí ČZU Praha a Ústí nad Labem. Kvůli identifikaci modelu je referenčnímu prvku (kterým je Brno) přiřazena hodnota odhadu 0, proto je nulová i směrodatná odchylka a p-hodnota. Oproti MODELu 1, viz 4.4, je větší rozdíl v odhadech koeficientů u nejlepších a nejhorších týmů, např. nejlepší tým soutěže Ostrava má v tomto modelu odhadnutou sílu 1.7121 (v MODELu 1 má 1.5822), zatímco nejhorší tým Benátky -1.1408 (-0.7903). To je způsobeno rovněž přítomností efektu domácí výhody v modelu. Hodnota koeficientu domácí výhody je přibližně 0.98, tj. poměr šancí na vítězství týmu v domácím prostředí a vítězství týmu bez této výhody je $\exp 0.98 = 2.66$. Tedy šance na vítězství se přechodem na domácí hřiště zvýší více než 2.5×, což je zjevně velká výhoda. Tento fakt ilustrujeme i na obrázku 4.1, na kterém jsou graficky znázorněny pravděpodobnosti výhry některých dvojic týmů v závislosti na místě konání zápasu. Z levého obrázku vidíme, že pravděpodobnost vítězství velkého favorita se na různých hřištích o tolik nemění - pravděpodobnost výhry Ostravy nad Benátkami nad Jizerou je na domácím hřišti 98%, na neutrální půdě je to 95%, a pokud hrají doma Benátky, pak Ostrava vyhraje s pravděpodobností 87%. Větší zisk konáním zápasu na domácí půdě získá slabší tým - v případě Benátek je to nárůst 8% na výhru oproti hře na neutrálním hřišti, u Ostravy se jedná o pouhá 3%. Větší rozdíly nastanou při střetu dvou vyrovnanějších týmů, jak ukazuje obrázek vpravo. Na neutrální půdě je pravděpodobnost výhry Českých Budějovic a Opavy podobná - 57%, resp 43%. Jestliže se ovšem hraje v největším městě Jižních Čech, pak České Budějovice vyhrají v 78% utkání, přejezdem na Moravu se naopak dramaticky zvyšuje pravděpodobnost výhry Opavských - 67%.

V další části již budeme pracovat s daty pouze pro sezonu 2007/08. Použijeme model s vysvětlujícími proměnnými, jejichž hodnoty jsou uvedeny v tabulce 4.3. Vyšli jsme z modelu:

$$\beta_i = \gamma_1 \texttt{Vaha}_i + \gamma_2 \texttt{Vyska}_i + \gamma_3 \texttt{Vek}_i + \gamma_4 \texttt{Divaci}_i, \tag{4.1}$$

kde navíc počítáme s výhodou domácího prostředí, tedy koeficienty β_i musí



Obrázek 4.1: *Grafické znázornění vlivu domácího prostředí na pravděpodobnosti vítězství v zápase vybraných týmů.*

splňovat MODEL 2. Některé proměnné v tomto modelu se ukázaly jako statisticky nesignifikantní a vyřadili jsme je - týkalo se to veličin `Divaci` a `Vek`. Finálním modelem tak je:

$$\beta_i = \gamma_1 \texttt{Vaha}_i + \gamma_2 \texttt{Vyska}_i \quad \text{MODEL 3}$$

Odhadnuté koeficienty pro dvě vysvětlující proměnné `Vaha` a `Vyska` a pro vliv domácího prostředí jsou uvedeny v tabulce 4.6. P-hodnoty u vysvětlujících proměnných `Vaha` a `Vyska` přesahují 0.05 (resp. 0.10), avšak obě proměnné jsme se rozhodli v modelu ponechat. Z koeficientů u vysvětlujících proměnných `Vaha` a `Vyska` se dá odvodit, že jeden kilogram průměrné váhy v týmu zvedne poměr šancí na vítězství přibližně $1.2\times$ ($\exp 0,2046 = 1,2270$), zatímco jeden centimetr průměrné výšky navýší poměr šancí na vítězství cca $1.8\times$ ($\exp 0,5841 = 1,7933$). Stejně jako v předchozím modelu i v tomto je přítomen signifikantní efekt domácího prostředí. Model s dosazenými koeficienty pro γ_1 a γ_2 je:

$$\beta_i = 0.2046\,\texttt{Vaha}_i + 0.5841\,\texttt{Vyska}_i \quad \text{MODEL 3}$$

Výsledný model má sice výše zmíněný tvar, ale ani ten není na první pohled ideální, a to z toho důvodu, že p-hodnoty u obou proměnných jsou vyšší než 0.05, což může naznačovat, že jsou statisticky nevýznamné. Z koeficientů u vysvětlujících proměnných `Vaha` a `Vyska` se dá odvodit, že jeden kilogram průměrné váhy v týmu zvedne poměr šancí na vítězství přibližně $1.2\times$ ($\exp 0,2046 = 1,2270$), zatímco jeden centimetr průměrné výšky navýší poměr šancí na vítězství cca $1.8\times$ ($\exp 0,5841 = 1,7933$). Volejbalovým

MODEL 2	odhad	sm. odchylka	p-hodnota	
Ostrava	1.7121	0.3769	5.54e-06	***
Liberec	1.4623	0.3658	6.39e-05	***
CB	1.1504	0.3553	0.00121	**
Kladno	1.1024	0.3584	0.00210	**
Opava	0.8574	0.3460	0.01320	*
CZU	0.1878	0.5118	0.71366	
Ústí	0.1606	0.3457	0.64216	
Brno	0	0	0	
Zlín	-0.1110	0.3478	0.74960	
Příbram	-0.2360	0.3529	0.50363	
Benátky nad Jizerou	-1.1408	0.4458	0.01049	*
domácí výhoda	0.9851	0.1378	8.93e-13	***

Tabulka 4.5: *Koeficienty $\beta_1, \ldots, \beta_{11}$ MODELu 2 odhadnuté v programu R. V prvním sloupci je uveden tým příslušný danému koeficientu.*

MODEL 3	odhad	sm. odchylka	p-hodnota	
Vyska	0.5841	0.3627	0.1072	
Vaha	0.2046	0.1082	0.0586	.
domácí výhoda	0.3946	0.1799	0.0283	*

Tabulka 4.6: *Koeficienty MODELu 3 odhadnuté v programu R.*

MODEL 3	síla	sm. odchylka
Liberec	1.0821746	0.4847494
Ostrava	0.8442760	0.4556119
Zlín	0.7505196	0.3406033
Kladno	0.7223157	0.3670516
Ústí	0.6742104	0.3025229
CZU	0.6683690	0.3002162
CB	0.6128207	0.2737998
Opava	0.6053810	0.2851518
Příbram	0.3154282	0.3529043
Brno	0.0000000	0.0000000

Tabulka 4.7: *Vypočtené síly týmů v MODELu 3.*

trenérům můžeme tedy doporučit, aby do svého týmu nabírali především vysoké a dobře rostlé svěřence.

V tabulce 4.7 jsou uvedeny vypočtené koeficienty $\beta_1, \ldots, \beta_{10}$ vyjadřující herní sílu týmů. Zobrazené odhady koeficientů se spočítaly tak, že se průměrná váha a výška týmu dosadila do výsledného MODELu 3, a následně se od ní odečetla takto vypočtená hodnota pro referenční prvek - Brno. Pokud si hodnoty koeficientů prohlédneme, zjistíme, že se pohybují v menším rozpětí než například odhady u prvních dvou modelů. Tedy MODEL 3 naznačuje velkou vyrovnanost napříč celou ligou, což ovšem příliš neodpovídá reálné tabulce, viz 4.2. Z toho důvodu si myslíme, že je náš model nedostačující, a pro lepší přiblížení se skutečnosti by bylo vhodné zahrnout do modelu další vysvětlující proměnné (např. rozpočet družstev).

Závěr

V této bakalářské práci jsme se zabývali Bradley-Terryho modelem a práci s ním ve statistickém softwaru `R`. V první kapitole jsme popsali teoretické základy nutné k definování Bradley-Terryho modelu, zejména jsme se věnovali zobecněnému lineárnímu modelu (GLM) a logistické regresi, z níž Bradley-Terryho model vychází.

V další kapitole jsme pak definovali základní Bradley-Terryho model. Protože ten na většinu situací pro párové porovnávání z reálného života nestačí, popsali jsme dále některé jeho modifikace. Důležitou pro modelování výsledků sportovních utkání je například Bradley-Terryho model zahrnující remízy nebo Bradley-Terryho model s efektem domácího prostředí. Na závěr této kapitoly jsme čtenáře stručně uvedli do problematiky odhadování parametrů Bradley-Terryho modelu. Parametry modelu nelze odhadnout přímo analyticky, obvykle se postupuje numerickým řešením optimalizačních úloh spojených s negativně logaritmickou funkcí. Použitelné jsou například tzv. MM algoritmy.

Ve třetí kapitole jsme se věnovali práci s Bradley-Terryho modelem v programu `R`, a to prostřednictvím předprogramovaného balíčku `BradleyTerry`. Funkce tohoto balíčku jsme ilustrovali na práci s datovým souborem vzájemných duelů osmi nejlepších tenistů světa. Pro tuto naši množinu hráčů jsme například zjistili, že levorucí mají v utkání s pravorukým hráčem větší šanci na vítězství.

Čtvrtá kapitola obsahuje rozbor výsledků sportovních utkání. Z datového souboru celkem 405 utkání volejbalových týmů během tří sezon jsme odhadli základní Bradley-Terryho model, a následně i dopad domácího prostředí na výsledek utkání. Zjistili jsme, že odhadnuté síly týmů dobře odpovídají skutečnému postavení v tabulce a že efekt domácího prostředí je ve volejbalu poměrně značný - šance na vítězství se týmu přechodem na domácí hřiště zvedne 2.66krát. Nakonec jsme zadefinovali model s exogenními proměnnými charakterizujícími jednotlivá mužstva a odhadli v něm vliv těchto charakteristik na výkon týmu. Zjistili jsme, že z námi zvolených vysvětlujících proměnných mají vliv průměrná výška a váha mužstva. Tyto proměnné však měly vyšší p-hodnoty a síly týmů vypočtené z konečného modelu příliš nekorespondovaly se skutečností a s naším očekáváním. Proto

musíme konstatovat, že na počátku našeho výzkumu jsme nezvolili dostatečný počet vysvětlujících proměnných a pro zpřesnění modelu by bylo třeba získat data o více charakteristikách volejbalových týmů.

Literatura

[1] A. Agresti: *Categorical Data Analysis*, Wiley, New York, 1990.

[2] A. Agresti: *An Introduction to Categorical Data Analysis*, Wiley, New York, 1996.

[3] R.A. Bradley a M.E. Terry: Rank analysis of incomplete block designs, I. the method of paired comparisons. *Biometrika*, **39**, 324—345, 1952.

[4] R. R. Davidson: On Extending the Bradley-Terry Model to Accommodate Ties in Paired Comparison Experiments, *Journal of the American Statistical Association*, **65**, 317—328, 1970.

[5] D. Firth : Bias reduction of maximum likelihood estimates, *Biometrika*, **80**, 27—38, 1993.

[6] D. Firth : Bradley-Terry Models, in R, *Journal of Statistical Software*, **12**, 1—12, 2005.

[7] T.K. Huang, R.C. Weng, C.J. Lin: Generalized Bradley-Terry Models and Multi-class Probability Estimates, *Journal of Machine Learning Research*, **7**, 85—115, 2006.

[8] D. R. Hunter: MM algorithms for generalized Bradley-Terry models, *The Annals of Statistics*, **32**, 386—408, 2004.

[9] R. D. Luce: *Individual Choice Behavior*, Wiley, New York, 1959.

[10] P. V. Rao a L. L. Kupper: Ties in Paired-Comparison Experiments: A Generalizition of the Bradley-Terry Model, *Journal of the American Statistical Association*, **62**, 194—204, 1967.

[11] H. Turner a D. Firth: *Bradley-Terry models in R: The BradleyTerry2 package*, http://cran.r-project.org/web/packages/BradleyTerry2/vignettes/BradleyTerry.pdf.

[12] E. Zermelo: Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung, *Math. Z.*, **29**, 436—460, 1929.

[13] K. Zvára: *Regrese*, Matfyzpress, Praha, 2008.

[14] *R 2.9.0.* R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

[15] Atpworldtour.com [online]. 2010 [cit. 2010-07-25]. Atpworldtour.com. Dostupné z WWW: <www.atpworldtour.com>.

[16] Cvf.cz [online]. 2010 [cit. 2010-07-31]. Cvf.cz. Dostupné z WWW: <www.cvf.cz>.

[17] Idnes.cz [online]. 2008 [cit. 2010-07-20]. Volejbal.idnes.cz. Dostupné z WWW: <www.idnes.cz>.